Face-Guided Sentiment Boundary Enhancement for Weakly-Supervised Temporal Sentiment Localization

Supplementary Material

Table 8. Ablation study of BSPG parameters (confidence decay factor β and smoothing width w) on mAP performance.

BSPG	w	5	6	7	8	9
	mAP (%)	19.22	19.77	**21.45**	19.66	19.26
	β	0.1	0.3	0.5	0.6	0.7
	mAP (%)	19.64	18.88	21.29	**21.45**	21.20

1. Hyperparameter Ablation

Ablation on BSPG Parameters. Our BSPG improves the reliability of pseudo-labels under point-level supervision through step-by-step smoothing. This reduces temporal jitter and mitigates boundary discontinuities (Fig. 2(c), Eq. 7). We ablate two key hyperparameters: the smoothing width w and the confidence decay factor β. As shown in Table 8, performance peaks at $w = 7$ (mAP: 21.45%), indicating that moderate neighborhood expansion effectively aggregates supervisory cues. Smaller windows (w = 5, mAP: 19.22%) under-smooth labels, whereas larger ones (w = 9, mAP: 19.26%) over-smooth and blur temporal boundaries. For the decay factor, optimal performance is achieved at $\beta = 0.6$. Lower values (β = 0.1, mAP: 19.64%) decay too quickly, limiting contextual information, whereas higher values (β = 0.7, mAP: 21.20%) overconfidently propagate pseudo-labels, introducing spurious activations. Overall, tuning both w and β improves pseudo-label quality and temporal localization.

Ablation on PSSC Parameters. Our PSSC enhances emotion boundary localization by leveraging the semantic space across temporal sequences, as illustrated in Fig. 2(c). The ablation study on the sample quantity in Eq. 5 is conducted based on $K = \lfloor T/k \rfloor$, where T denotes the total number of frames. This design accounts for significant temporal variations across input videos. As shown in Fig. 5, the hyperparameter k has a substantial impact on performance, with mAP showing a nonlinear trend. The best result (21.5% mAP) occurs at $k = 8$, indicating a balance between emotional coverage and noise suppression. A moderate sample count promotes the optimization of $\mathcal{L}_{\text{sc}}$ by aggregating discriminative emotional patterns and capturing the spatial distribution of frame-level F_{mix}. In contrast, $k = 5$ yields insufficient cues, while $k = 12$ introduces redundancy.

Ablation on Loss Weights. We explore the impact of loss weights on the experiment in terms of mAP. As shown in Fig. 6, the blue curve indicates the variation of λ_1 when $\lambda_2 = 0.05$, while the green curve shows the variation of λ_2 when $\lambda_1 = 1$. The mAP reaches its peak at $(\lambda_1, \lambda_2) = (1, 0.05)$, highlighting the importance of properly balancing the two loss terms. Specifically, $\mathcal{L}_{\text{frame}}$ improves the accuracy of emotion localization, whereas $\mathcal{L}_{\text{sc}}$ helps align frame-level emotion segments. A small λ_1 results in insufficient boundary supervision, while an overly large λ_2 may over-constrain the alignment, leading to blurred segment boundaries and loss of fine-grained emotional cues.

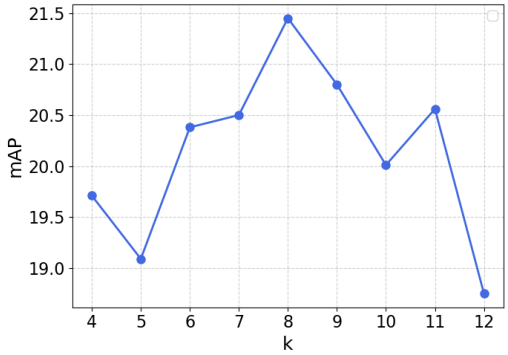

Figure 5. Ablation study of Top-K parameter k in PSSC on mAP performance.

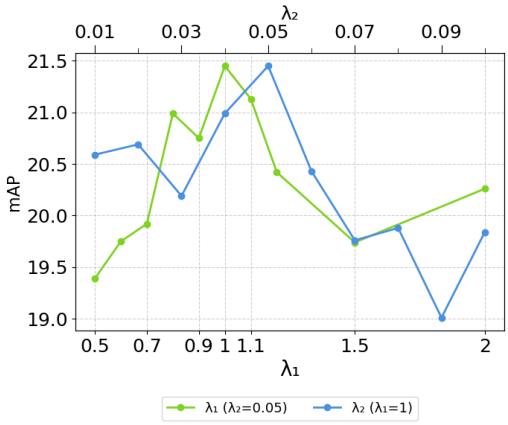

Figure 6. Ablation study of loss function weights λ_1 and λ_2 on mAP performance.

2. Additional Results

Error Analysis. To gain deeper insight, we follow the error diagnosis protocol of [2] to break down the model outputs on the TSL300 dataset. As in Fig. 7, a key observa-

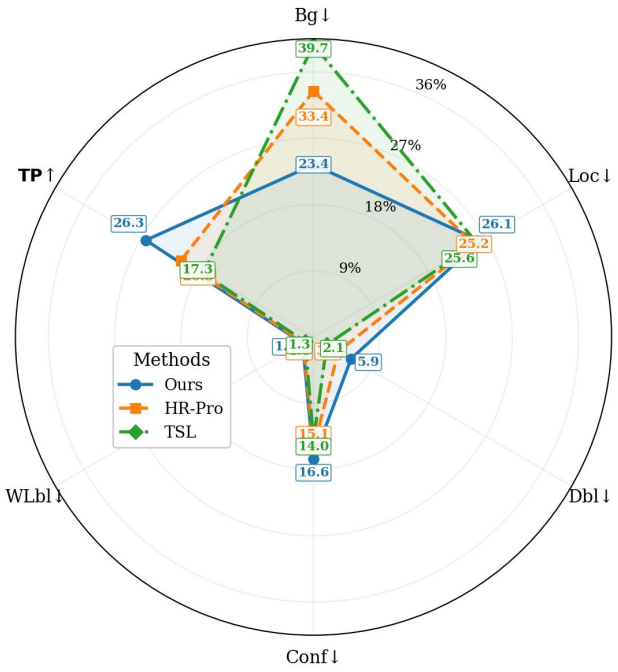

Figure 7. Proportion of proposal type for all prediction results of models on TSL300. Error categories: Bg (background), Loc (localization), Dbl (double detection), Conf (confusion), WLbl (wrong label), and TP (true positive).

tion is the substantial reduction in background errors, with a 10.0% decrease compared to HR-Pro (33.4% → 23.4%) and a 16.3% decrease compared to TSL (39.7% → 23.4%). This is accompanied by a notable increase in true positives (TP: 26.3%). These results strongly suggest that our approach effectively reduces false positives arising from background confusion. The fact that localization error (Loc) remains the dominant error type also indicates potential for further improvement in overall detection accuracy (mAP), pointing to future work on refining temporal boundaries.

PSSC Distance Metric Ablation. To assess the impact of different distance metrics on the identification of semantically relevant positive samples, we conduct an ablation study on the similarity function used in Eq. (4) of the PSSC module. Table 10 reports the mAP results under multiple IoU thresholds. Among the compared metrics, **Cosine similarity (Ours)** achieves the highest average mAP of 21.45%. In comparison, the L2 distance, the L1 distance and the dot product achieve lower scores of 20. 29%, 20. 05% and 21. 00%, respectively. This corresponds to absolute improvements of 1.16%, 1.40%, and 0.45%. Under low IoU thresholds (0.1, 0.15, 0.2), where spatial constraints are loose and semantic disentanglement is more difficult, Cosine still achieves the best performance, with mAP of 29. 31%, 25. 47%, and 22. 49%. These results demonstrate the effectiveness of cosine similarity in capturing semantic relationships between feature vectors.

Table 9. Performance comparison of different modality combinations on the TSL task. A: Audio, V: Visual, F: Local Face. Metrics are reported as mAP@IoU (%) across various thresholds and average mAP.

Modality	**mAP@IoU (%)** 0.1	0.15	0.2	0.25	0.3	**Avg mAP**
Single Modality						
A (Audio)	18.98	15.47	12.96	9.75	7.09	12.85
V (Visual)	19.27	16.14	12.06	9.14	6.45	12.61
F (Local Face)	22.11	15.97	11.33	8.40	6.62	12.88
Bimodal Combinations						
A + F	29.41	24.01	18.70	13.97	9.95	19.21
A + V	25.91	23.70	20.08	15.85	11.53	19.41
V + F	26.20	22.48	17.41	14.53	11.30	18.38
Trimodal Combination						
A + V + F	**29.31**	**25.47**	**22.49**	**16.76**	**13.24**	**21.45**

Table 10. Ablation of feature distance metrics for Point-aware Sentiment Semantics Contrast modeling.

Method	**mAP@IoU (%)** 0.1	0.15	0.2	0.25	0.3	**Avg mAP**
L2	28.06	24.40	20.98	16.49	11.55	20.29
L1	26.53	23.66	20.70	16.36	12.97	20.05
Dot	27.23	24.90	21.21	**17.42**	**14.22**	21.00
Cosine(Ours)	**29.31**	**25.47**	**22.49**	16.76	13.24	**21.45**

Modality Combination Analysis. We conduct an ablation study to examine the impact of the modality information, as shown in Table 9, which reports the performance of different modality combinations. Among the unimodal inputs, the face embedding (F) achieves the best performance (mAP: 12.88%), slightly outperforming the audio (A) and visual (V) features. This suggests that each modality provides complementary yet insufficient information when used in isolation. Bimodal combinations such as A+V and A+F yield significantly improved results (mAP: 19.41%, representing an increase of 6.56%; and 19.21%, with an increase of 6.36%, respectively). These results indicate strong complementarity between audio and visual/facial cues. Meanwhile, the relatively lower performance of V+F (18.38%) may be attributed to redundancy between holistic and local visual features. Trimodal fusion (A+V+F) further boosts the performance to 21.45% mAP, demonstrating the advantage of integrating heterogeneous modalities.